Wrangle Report

#WeRateDogs

The goal for this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. I will be using Python and its libraries to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

The dataset that I will be wrangling is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Gathering Data

1. Enhanced Twitter Archive: contains basic tweet data for all 5000+ of their tweets.

2. Image Predictions File: contains a table full of image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

3. Additional Data via the Twitter API: contains retweet count and favorite count which are two of the notable column .

Assessing Data

This step allows us to identify quality and tidiness issues.

Quality: issues with content.

- Low quality data is also known as dirty data such as missing, invalid, inaccurate and inconsistent data.

Tidiness: issues with structure that prevent easy analysis.

- Untidy data is also known as messy data.

Two types of assessment are done to identify the issues:

Visual assessment: scrolling through the data
Programmatic assessment: using code to view specific portions and summaries of the data

The following is the list of issues that have been observed and will proceed to clean some of them in the next section:

Quality

Twitter Archive Table

- Many variables need to convert to the right datatype (timestamp, source, doggo, floofer, pupper and puppo).
- There are missing name and some of the name fields contains prepositions (e.g. 'a', 'actually', 'all', etc).
- The column headers are not descriptive .
- Data contains retweets (ie. rows where retweeted_status_id and retweeted_status_user_id have a number instead of NaN).
- Delete the unneccessry colummns.
- Split date to day,moth,year and time columns.

Image Prediction Table

- Variable need to convert to the right datatype (img_num).
- Consolidate the numerous image prediction columns to new column and delete those unwanted column(i.e,that columns have no impact on analysis)

Twitter API Table

- Rename the id in twitter_api_df to twitter_id for joining the 3 dataframes (i.e, using the common column ' twitter_id' for merging 3 datafarmes)

Tidiness

- Merge all the 3 DataFrames.
- dog stage contains in 4 different columns (doggo, floofer, pupper, and puppo)

Cleaning Data

The data have been cleaned using the programmatic methods. They are shown under the define, code and test format.

- Define: definition or issue to clean of fix.
- Code: the code use to clean or fix the issue.
- Test: to assure that the cleaning operations worked correctly.

Conclusion

Through the data wrangling process, we managed to fix a number of problems and cleaned the data for further analysis to be done.